Implementation of machine learning into routine clinical microbiology

Adrian Egli, MD PhD
Head Clinical Bacteriology & Mycology, University Hospital Basel
Research Group Leader Applied Microbiology Research, University of Basel
Agenda

- Why we need machine learning?
  - Increasing complexity of microbiological data
  - Interface of clinics and laboratory

- How to integrate data driven microbiology in routine diagnostics
  - Requirements
  - What is already here?
  - Outlook: ontologies and data warehouses
Status quo of modern clinical microbiology

- **Culture**
  - Automation\(^1,2\): Throughput
  - Standardized agar plates\(^3\)
  - Deliverables: Growth, semi-quantitative

- **MALDI-TOF MS\(^4\)**
  - Faster identification\(^5\)
  - Lower costs \(^6,7\)
  - Deliverable: Precise species ID

---

\(^7\) Tan KE, Ellis BC et al. JCM 2012; 50:3301-8
Status quo of modern clinical microbiology

- **Molecular diagnostics**
  - Culture independent
  - Identification and resistance testing e.g. *M. tuberculosis* and rifampicin resistance\(^1,2\)
  - Point of care testing\(^3\)

- **Change in diagnostic focus**
  - From single targets towards panel PCRs\(^4,5\)
  - Mixing of traditions: virology & bacteriology

- **Deliverable**: Gene detection, quantification

---

5 Cybulski RJ Jr, Bateman AC et al. CID, 67:1688-1696
Current communication optimized for status quo

Clinical Microbiology
- Simple results delivered
- Binary: yes/no
- Semi-quantitative CFU/mL
- Quantitative: GEq/mL
- Results: short lists/categories
- Pathogen-centric

Physician
- Simple results expected
- Binary algorithms
- Patient-centric

→ More complex results & higher quantity of results
→ Challenge in (i) interpretation and (ii) communication
Example 1: Broad panel PCRs are challenging...

- PCR technology -> high sensitivity
- Detection of **20+ pathogens**
- Various **syndromic panels**:
  - Meningitis & encephalitis
  - Upper & lower respiratory tract infection
  - Bacteremia
  - Diarrhea
  - Sexually transmitted infections
- Problems:
  - Costly
  - Who benefits?
What is the consequence of panel assays?

- **Paradigm change**: “Think of what you want” towards “know the gaps!”
- No longer pathogen-specific thinking, but **syndromes**!

- **Evidence**: few clinical impact studies available\(^1,2,3\)!

  Often microbiological endpoint e.g. sensitivity/specificity\(^4,5\)

---

What is the consequence of panel assays?

- **Paradigm change**: “Think of what you want” towards “know the gaps!”
- No longer pathogen-specific thinking but **syndromes**!

- **Evidence**: very few clinical impact studies available\(^1,2,3\)!
  Often microbiological endpoint e.g. sensitivity/specificity\(^4,5\)

- **Not an easy study design**, because…
  - State of the art diagnostics -> ethical conflict?

→ **Radical new study designs** driven by machine learning:
  e.g. reinforcement learning
  **Determination who benefits most from a test!**

---

Example 2: Metagenomic from blood culture

Goal: (i) Metagenomic data in 3-4h available for routine diagnostics
(ii) Species identification and detection of resistance genes

Bonfiglio F et al. unpublished
Identification of species without problems

**E. coli**

**K. pneumoniae**

**S. aureus**

Bonfiglio F et al. unpublished
Interpretation of genetic resistance needs work…

→ Identification is no problem
→ Resistance testing is challenging
  - Gene detection ≠ phenotype
  - Quality of databases
  - Interpretation of complex information

Bonfiglio F et al. unpublished
What is the consequence of metagenomics?

- Once technical challenges are solved... e.g. contamination, standards\textsuperscript{1,2,3}
- Delivery: \textit{relative ratios of bacteria}! What is a pathogen?
- Evidence: \textit{very few clinical impact studies} yet available\textsuperscript{4,5}!
- Complexity of data is almost limitless...\textsuperscript{3}

→ Microorganism identification
→ Antibiotic resistance prediction
→ Detection of virulence determinants
→ Antiviral resistance predication
→ Microbiome analysis
→ Transcriptomics
→ Oncology applications

1 Thoendel M, Jeraldo P et al. JCM 2017; Mulcahy-O’Grady H et Workentine ML Front Immunol 2016;
More complex results require a longer “brain-to-brain” time

Laboratory information system (LIS)
Clinical information system (CIS)
+ visualize the data

Clinical Microbiology
- Complex results delivered
- Quantitative: GEq/mL
- “-omics”
- Results: Big data
- Is it really a pathogens?

Physician
- How to handle complex data?
- What does the data mean?
- No longer: “yes or no”
- “Hypothesis free”
- What can I do for the patient
- What is a disease?

Multiple partners in communication
Microbiologist & Bioinformatician & Data scientist
“Microbiolgy data board”
Consequences of technological development

- Explosion of knowledge in microbiology
- High specialization: from pathogens to technology (to data…)
- Do we really want machine learning as the next revolution?
Consequences of technological development

- Explosion of knowledge in microbiology
- High specialization: from pathogens to technology (to data…)

You will need machine learning because of resulting challenges:

- How to manage, interpret, and communicate big data?
- What is important data and what is not?
- Generate impact of data driven microbiology for the patient!
Complexity and quantity of information will further increase
→ How to handle big data in clinical workflows?
→ How to generate impact for the patient?

→ Machine learning may help to sort and understand data
Key steps to implementation of machine learning

(i) Requirement analysis & process
- Workflow in the lab: Pre-analytic, analytic and post-analytic
- What are the critical nodes/interfaces?
- What results is critical: content, delivery time, receiver?
Key steps to implementation of machine learning

(i) Requirement analysis & process

(ii) Autonomous commercial systems
- Investment in information technology and teams
e.g. LIS, alerting system, Apps, visualization tools, IT specialists
- Machine learning based diagnostics
e.g. microscopy, culture morphology
Microscopy: Gram stain classification

Smith KP, et al. JCM 2018

- Automated image acquisition with MetaFer platform, 40x objective

- Training: 180 slides → 25’488 images → 100’213 crops

Typical image
Microscopy: Gram stain classification

- Convolutional neural network: Classification accuracy of 94.9%
- ROC robust ability to differentiate between categories with AUC >0.98

- Validation dataset: 189 new slides → Sensitivity and specificity

**TABLE 1** Confusion matrix of whole-slide classification results

<table>
<thead>
<tr>
<th>Human classification</th>
<th>Gram negative</th>
<th>Gram-positive pairs or chains</th>
<th>Gram-positive clusters</th>
<th>Background</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gram negative</td>
<td>51</td>
<td>1</td>
<td>0</td>
<td>17</td>
</tr>
<tr>
<td>Gram-positive pairs or chains</td>
<td>3</td>
<td>27</td>
<td>6</td>
<td>4</td>
</tr>
<tr>
<td>Gram-positive clusters</td>
<td>1</td>
<td>1</td>
<td>70</td>
<td>8</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>% sensitivity (CI)α</th>
<th>% specificity (CI)α</th>
</tr>
</thead>
<tbody>
<tr>
<td>98.1 (94.3–100)</td>
<td>96.3 (93.7–98.9)</td>
</tr>
<tr>
<td>75.0 (60.9–89.0)</td>
<td>98.4 (90.8–100)</td>
</tr>
<tr>
<td>97.2 (93.4–100)</td>
<td>93.2 (89.7–96.6)</td>
</tr>
</tbody>
</table>

*Data were determined based on slides where bacteria were detected. CI, 95% confidence interval.*
Machine learning in microscopy

- **Concept:** Recognition of patterns
- Large datasets necessary → fast saturation of accuracy in training set

- Various types of staining in the lab: e.g. Gram, Acridine orange, Ziehl Neelsen, Auramin Rhodamin

- **Quality:** very critical e.g. manual vs. automated staining, background

- **Studies:** Only one study¹

- Optimization of studies for sensitivity and specificity:
  - In positive blood cultures: specificity is more important
  - In non-precultured material e.g. CSF: sensitivity is more important

¹ Smith KP, et al. JCM 2018
Digital plate reading: detection and ID

- **Detection of colonies**: positive vs. negative plates
  - BD Kiestra system with Sensitivity 97.1%; Specificity 93.6%\(^1\)
  - APAS >99% sensitivity and specificity on BA and MacConkey agar\(^2\)

- **Identification of bacterial species**
  - BD Kiestra identification of e.g. *S. aureus*, KECS, *Enterococcus* spp. with specificity of 93.8% - 99.2%\(^1\)
  - *E. coli* on APAS >99%\(^2\)

Digital plate reading: through put

Negative plates

Positive plates

Continuous improvement of pattern recognition
- Present/absent
- Identification

PhenoMATRIX
(Copan/bioMérieux)
Machine learning in plate reading

- Detection of bacteria on agar plates\(^1,\text{2}\)
  - Growth: Yes/No
  - Identification: Species
  - Screening MRSA\(^3\) or VRE\(^4\)
    - High sensitivity (100\%) and specificity (>90\%) with PhenoMATRIX\(^3,\text{4}\)

- Comparing machine learning studies is difficult:
  - Heterogenous samples
  - Training dataset differ
  - Algorithms and methods used are different

---

Key steps to implementation of machine learning

(i) Requirement analysis & process optimization

(ii) Autonomous commercial systems

(iii) Research and development for future applications

- Common language: ontologies
- Data warehouses to collect information
- Prospective (!) studies to explore the impact of machine learning
R&D requirement 1: Ontologies

- **Problem:** Medical data can be determined and defined in various ways e.g. laboratory values MICs: Basel ≠ Amsterdam

- **What ontologies provides?**
  - Common vocabulary
  - Definitions
  - Structure for data
  - Results in comparability and interoperability

- **Examples:**

![SNOMED CT](image1)

![LOINC](image2)

![IRIDA](image3)
R&D requirement 2: Clinical data warehouse

- Data warehouse\(^1,2\): repository of historical granular patient-centric data for reporting and analysis. Facilitates data access by having data in one place.

---

The world of machine learning just started…

- **Types of learning algorithms**
  - Supervised and semi-supervised learning e.g. infection on admission\(^1\), inter-species relationships\(^2\)
  - Unsupervised learning e.g. transposon insertion sites\(^3\)
  - Reinforcement learning e.g. innate immune response to infection\(^4\)

- **Processes and techniques**
  - Feature learning e.g. immunoprofiling of latent tuberculosis\(^5\)
  - Sparse dictionary learning e.g. EEG\(^6\)
  - Anomaly detection e.g. labelling of x-rays\(^7\)
  - Decision trees e.g. Urinary tract infection\(^8\)
  - Association rules e.g. identification of influenza host range\(^9\)

- **Models**
  - Artificial neural networks aka deep learning e.g. detecting MDR TB\(^{10}\), forecasting norovirus\(^{11}\)
  - Support vector machines e.g. heart rate and sepsis\(^{12}\)
  - Bayesian networks e.g. malaria control\(^{13}\)
  - Genetic algorithms e.g. Shigatoxin outcome\(^{14}\)

---

Example: Digital biomarkers in sepsis

- Reinforcement learning for decision-making of sepsis treatment

- 48 variables including vital signs, laboratory values, fluids and vasopressor received
- Retrospective analysis of data

⇒ Value of the AI selected treatment is on average reliably higher than human clinician

Outlook: Where companies should invest…

- **Prospective (!) validation for clinical application**
  - Validate a DNN in silico
  - Clinical validation in real-world medicine
  - Implementation in healthcare

- **Supply chain management** in the lab
  - Data: Integration of test frequencies, ward data e.g. bed occupancy, season
  - Personnel planning e.g. in core labs, diversity of experts
  - Reagents demand, automated ordering

- **Diagnostic stewardship**
  - Data: Case specific information e.g. immunosuppressed, travel history
  - When to use, which (expensive) test? E.g. broad panel PCR

- **Identification of species**
  - Data: Photo from plate e.g. culture morphology, growth pattern
  - Culture morphology

---

Topol EJ Nat Med 2019
Outlook: Where companies should invest…

- **Antibiotic resistance**
  - Data: Expert systems, integration of metadata e.g. stay abroad, colonization
  - Interpretation of breakpoints
  - Recognition of MDR pathogen

- **Hospital epidemiology surveillance**
  - Data: Same species ID and resistance profile, spatiotemporal links
  - Potential transmission flagging

- **Public health surveillance**
  - Data: Surveillance platform for whole genome sequencing¹, spatiotemporal data, vaccine data
  - Flagging of epidemic and predicting dynamics
  - Identification of sources

¹ [www.spsp.ch](http://www.spsp.ch)
Validation of algorithms is critical

- Problems: (i) **high quality big data** and (ii) **team willing to share data**

- **Large consortia** to collect large high quality datasets
  - Example: SPHN/PHRT
  - Goal: Connect all Swiss Universities and University Hospitals

- Driver Project: Personalized Swiss Sepsis Study\(^1\)

- **Prospective validation studies** of AI algorithms\(^2,3\)
  - Impact for patient management, outcomes
  - Impact on workflows in laboratories e.g. turn around times, costs

---

1 www.sphn.ch; 2 Topol EJ Nat. Med 2019; 3 Rajkomar A, Dean J et Kohane I NEJM 2019
Critical aspects of machine learning

- “Garbage in -> garbage out”
- No replacement of humans(!)

Handling of information with support algorithms
- Is the AI given access to all variables that influence decision making?
- Publication bias! Many failures are not published
- Will the AI behave prospectively as intended?

Implementation of machine learning and data driven microbiology

→ Critical requirement analysis
→ Only few commercial available systems on the market
→ Prospective validation is required

SUMMARY PART II
Take home messages

Technological revolutions

Data driven microbiology
- Complexity & quantity of data
- “Hypothesis free”

Increase of knowledge
- Data management has to change
- Adapt communication

Machine learning will support, but will not replace
Acknowledgments & Collaborations

Research team @USB/DBM

Dr. H. Seth-Smith
Deputy

Dr. D. Wüthrich
Bioinformatics

Dr. F. Bonfiglio
Bioinformatics

A. Cuénod
PhD Student

J. Linnik
PhD Student

M. Syedbasha
PhD Student

D. Lang
Technician

J. Reist
Technician

R2DR

Various projects
Stefano Bassetti
Manuel Battegay
Nina Khanna
Michael Osthoff
Sarah Tschudin Sutter
Andreas Widmer

Various projects:
Karsten Borgwardt
Sebastian Bonhoeffer
Wolf-Dietrich Hardt
Alex Hall
Tanja Stadler
Jörg Stelling

SPHN/PHRT: driver project on sepsis
Network from all University Hospitals:
Basel, Bern, Geneva, Lausanne, Zurich
Infectious Diseases, ICUs and Microbiology
Digitalization and Infectious Diseases: Improving patient outcome in the age of big data

SAVE THE DATE
19-22 January 2020, Basel, Switzerland
website: Digital-ID2020.ch
Thank you for your attention!

Adrian Egli, MD PhD, FAMH

Head Clinical Bacteriology
University Hospital Basel

Research group leader
Applied Microbiology Research
University of Basel

Phone: +41 61 556 5749
Email: adrian.egli@usb.ch

Digitalization and Infectious Diseases
SAVE THE DATE
19-22 January 2020, Basel, Switzerland
website: Digital-ID2020.ch